Using Background Knowledge to Construct Bayesian Classifiers for Data-Poor Domains

Marcel van Gerven
Institute for Computing and Information Sciences
University of Nijmegen, Toernooiveld 1
6525 ED Nijmegen, The Netherlands
E-mail: marcelge@cs.kun.nl

Peter Lucas
Institute for Computing and Information Sciences
University of Nijmegen, Toernooiveld 1
6525 ED Nijmegen, The Netherlands
E-mail: peterl@cs.kun.nl

Abstract

The development of Bayesian classifiers is frequently accomplished by means of algorithms which are highly data-driven. Often, however, sufficient data are not available, which may be compensated for by eliciting background knowledge from experts. This paper explores the trade-offs between modelling using background knowledge from domain experts and machine learning using a small clinical dataset in the context of Bayesian classifiers. We utilised background knowledge to improve Bayesian classifier performance, both in terms of classification accuracy and in terms of modelling the structure of the underlying joint probability distribution. Relative differences between models of differing structural complexity, which were learnt using varying amounts of background knowledge, are explored. It is shown that the use of partial background knowledge may significantly improve the quality of the resulting classifiers.

1 Introduction

Again and again, Bayesian classifiers have proved to be a robust machine learning technique in the presence of sufficient amounts of data [3, 8, 5]. The heavy reliance of their construction algorithms on available data is, however, not always justified, as there are many domains in which this availability is limited. For instance, in the medical domain, more than 90% of medical disorders have a sporadic occurrence and, therefore, even clinical research datasets may only include data of a hundred to a few hundred patients. Clearly, in such cases there is a role for human domain knowledge to compensate for the limited availability of data, which then may act as background knowledge to a learning algorithm.

Even if the exploitation of background knowledge seems difficult to avoid in such data-poor domains, there is a question as to the form of this background knowledge. In the context of Bayesian classifiers, where the aim is to learn a probability distribution that is then used for classification purposes, representing background knowledge as a Bayesian network seems to have at least the appeal that it can easily be transferred to a Bayesian classifier. We call Bayesian networks that offer a task-neutral representation of statistical relations in a domain *declarative* Bayesian networks. Often, declarative Bayesian networks may be given a causal interpretation.

The construction of declarative Bayesian networks is a difficult undertaking; experts have to state perfectly all the dependencies, independencies and conditional probability distributions associated with a given domain. Since this is a very time-consuming task and an instantiation of the infamous *knowledge acquisition bottleneck*, we will investigate how background knowledge of different degrees of completeness influences the quality of the resulting classifiers built from this knowledge. We will refer to this form of incomplete and fragmentary knowledge as *partial background knowledge.*

We will use so-called *forest-augmented naive classifiers* in order to assess the performance of Bayesian classifiers of different degrees of structural complexity. Both the naive and the tree-augmented naive classifier are limiting cases of this type of Bayesian network [5]. Since Bayesian classifiers ultimately represent a joint probability distribution, we are not only interested in classifier performance, but also in the quality of the learnt probability distributions. The aim of this article is to gain insight into the quality of Bayesian classifiers when learnt from (partial) background knowledge instead of data.

2 Forest-augmented naive classifiers

2.1 Definition and construction

A *Bayesian network* $\mathcal{B}$ (also called belief network) is defined as a pair $\mathcal{B} = (G, P)$, where G is a directed, acyclic graph $G = (V(G), A(G))$, with a set of vertices $V(G) = \{X_1, \ldots, X_n\}$, representing a set of stochastic variables, and a set of arcs $A(G) \subseteq V(G) \times V(G)$, representing conditional and unconditional stochastic independences among the variables, modelled by the absence of arcs among vertices. Let $\pi_G(X_i)$ denote the conjunction of variables corresponding to the parents of X_i in G. On the variables in $V(G)$ is defined a joint probability distribution $P(X_1, \ldots, X_n)$, for which, as a consequence of the local Markov property, the following decomposition holds: $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \pi_G(X_i))$.

In order to systematically assess the performance of Bayesian classifiers with structures of varying complexity we introduce the *forest-augmented naive classifier*, or FAN classifier for short (Fig. 1). A FAN classifier is an extension of the naive classifier, where the topology of the resulting graph over the evidence variables $\mathcal{E} = \{E_1, \ldots, E_n\}$ is restricted to a forest of trees [5]. For each evidence

variable E_i there is at most one incoming arc allowed from $\mathcal{E} \setminus \{E_i\}$ and exactly one incoming arc from the class variable C.

The algorithm to construct FAN classifiers used in this paper is based on a modification of the algorithm to construct *tree-augmented naive* (TAN) classifiers by Friedman et al. [3] as described in [5], where the *class-conditional mutual information* (CMI)

$$I_P(E_i, E_j \mid C) = \sum_{E_i, E_j, C} P(E_i, E_j, C) \log \frac{P(E_i, E_j \mid C)}{P(E_i \mid C) P(E_j \mid C)} \tag{1}$$

is used to select succeeding arcs between evidence variables.

In our research, the joint probability distributions of the classifiers were learnt either from data using Bayesian updating with uniform Dirichlet priors or estimated from a declarative Bayesian network. We refer to classifiers of the first kind as *data-driven* classifiers (denoted by F_d) and to classifiers of the second kind as *model-driven* classifiers (denoted by F_m). We use F_k^n to refer to a type k FAN classifier containing n arcs. Note that F^n is equivalent to a naive classifier when $n = 0$ and equivalent to a TAN classifier when the arcs in F^n form a spanning tree over the evidence variables.

2.2 Estimating classifiers from background knowledge

The new approach studied in this article is to learn a Bayesian classifier's joint probability distribution not only from data, but alternatively to estimate it from a *declarative* Bayesian network. Declarative Bayesian networks may be viewed as the best approximation to the underlying probability distribution of the domain given the knowledge we have at our disposal. Learning FAN classifiers directly from a declarative model is accomplished as follows.

If we have a joint probability distribution $P(\mathcal{X}, \mathcal{E}, C)$ with $\mathcal{X} = \{X_1, \ldots, X_n\}$, evidence variables $\mathcal{E} = \{E_1, \ldots, E_m\}$ and class-variable C, underlying the declarative Bayesian network $\mathcal{B} = (G, P)$, then the following decomposition is associated with the Bayesian network:

$$P(\mathcal{X}, \mathcal{E}, C) = P(C \mid \pi_G(C)) \prod_{k=1}^{m} P(E_k \mid \pi_G(E_k)) \prod_{l=1}^{n} P(X_l \mid \pi_G(X_l)).$$

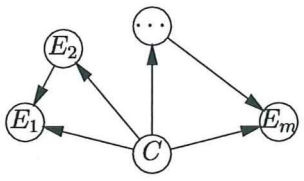

Figure 1: Forest-augmented naive (FAN) classifier. Note that both the naive classifier and the tree-augmented naive classifier are limiting cases of the forest-augmented naive classifier.

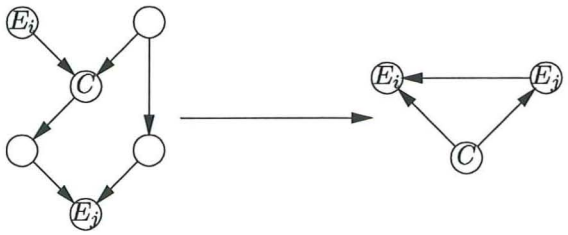

Figure 2: Declarative Bayesian network, used in computing the joint probability distributions for a three-vertex network, where $P(E_i, E_j, C) = P(E_i \mid E_j, C)P(E_j \mid C)P(C)$ and $P(E_i, E_j \mid C) = P(E_i \mid E_j, C)P(E_j \mid C)$.

The joint probability distribution underlying the FAN classifier $\mathcal{B}' = (G', P')$ with $V(G') = V(G)$ is defined as $P'(\mathcal{E}, C)$. The probability distribution P is used as a basis for the estimation of P', as follows:

$$P'(E_i \mid \rho(E_i), C) = \sum_{\gamma \in \sigma(\mathcal{X} \cup \mathcal{E} \setminus \{E_i\} \cup \rho(E_i))} P(E_i, \gamma \mid \rho(E_i), C) \qquad (2)$$

where $\sigma(\mathbf{V})$ denotes the set of configurations of the variables in $\mathbf{V}$ and

$$\rho(E_i) = \begin{cases} \{E_j\} & \text{if } \pi_{G'}(E_i) = \{E_j, C\} \\ \varnothing & \text{otherwise.} \end{cases}$$

The construction of FAN classifiers from the declarative model and the FAN construction algorithm amounts to estimating three-vertex networks of the form depicted in Fig. 2 using equation (2).

Since FAN classifiers may incorporate just a proper subset of the vertices in the declarative model, we are allowed to remove vertices which do not take part in the computation of the (conditional) probabilities $P(C)$, $P(E_j \mid C)$ and $P(E_i \mid E_j, C)$. Equation 2 does not take these irrelevant vertices explicitly into account, but standard techniques from the context of Bayesian inference exist to prune a declarative model prior to computing relevant probabilities [9].

2.3 Classifier evaluation

The performance of FAN classifiers may be determined by computing *zero-one loss*, where the value c^* of the class variable C with largest probability is taken: $c^* = \text{argmax}_c P(C = c \mid \mathcal{E})$.

A disadvantage of this straightforward method of comparing the quality of the classifiers is that the actual posterior probabilities are ignored. A more precise indication of the behaviour of Bayesian classifiers is obtained with the *logarithmic scoring rule* [2]. Let D be a dataset, $|D| = p$, $p \geq 0$. With each prediction generated by a Bayesian model for case $r_k \in D$, with actual class value c_k, we associated a score $S_k = -\log P(c_k \mid \mathcal{E})$, which can be interpreted formally as the entropy and has the informal meaning of a penalty. When the probability $P(c_k \mid \mathcal{E}) = 1$, then $S_k = 0$ (actually observing c_k generates no

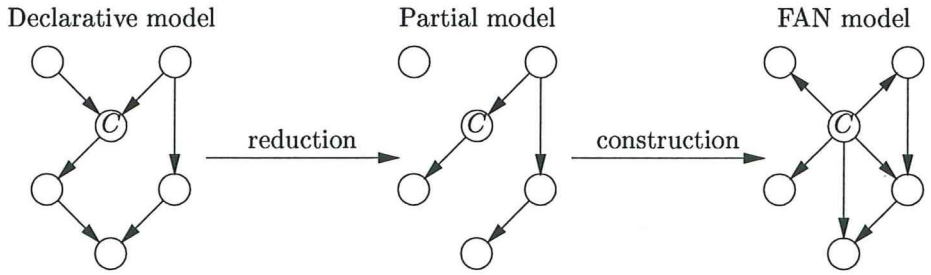

Figure 3: A declarative model is reduced to a partial model. Subsequently, FAN models are constructed from the partial model.

information); otherwise, $S_k > 0$. The total score for dataset D is now defined as the average of the individual scores $S = \frac{1}{p}\sum_{k=1}^{p} S_k$.

The logarithmic scoring rule is a rule which measures differences in probabilities for a class c_k given evidence $\mathcal{E}$. A global measure of the distance between two probability distributions P and Q is the *Kullback-Leibler* (KL) *divergence*:

$$\delta(P,Q) = \sum_{X} P(X) \log \frac{P(X)}{Q(X)}.$$

We have used the percentage of correctly classified cases computed using zero-one loss as our measure of classification accuracy, the logarithmic score to gain insight into the quality of the assigned probabilities for unseen cases and KL divergence as a means to gain insight into the quality of the joint probability distribution when comparing the declarative model with the other models.

2.4 Partial background knowledge

Declarative Bayesian networks are particularly useful to represent the background knowledge we have about a domain, but often this knowledge is incomplete. We define *partial background knowledge* as any form of knowledge which is incomplete relative to the total amount of background knowledge available. More formally, let $\mathcal{B} = (G, P)$ be a declarative model with joint probability distribution $P(X_1, \ldots, X_n)$, representing full knowledge of a domain. Let $\mathcal{B}' = (G', P')$ with $V(G') = V(G)$ be a Bayesian network with $P'(X_1, \ldots, X_n)$. $\mathcal{B}'$ is said to represent partial background knowledge if $0 < \delta(P, P') < \epsilon$ for small $\epsilon > 0$, where ϵ is the least upper-bound of $\delta(P, P')$ for an uninformed prior P'.

In this article we have focused on the incomplete specification of dependencies as our operationalisation of partial background knowledge, such that for a *partial model* $\mathcal{B}'$, $A(G') \subseteq A(G)$. The probability distribution P is used as a basis for the estimation of P', as follows:

$$P'(X_i \mid \pi_{G'}(X_i)) = \sum_{\gamma \in \sigma(\pi_G(X_i) \setminus \pi_{G'}(X_i))} P(X_i \mid \pi_{G'}(X_i), \gamma) P(\gamma \mid \pi_{G'}(X_i)). \quad (3)$$

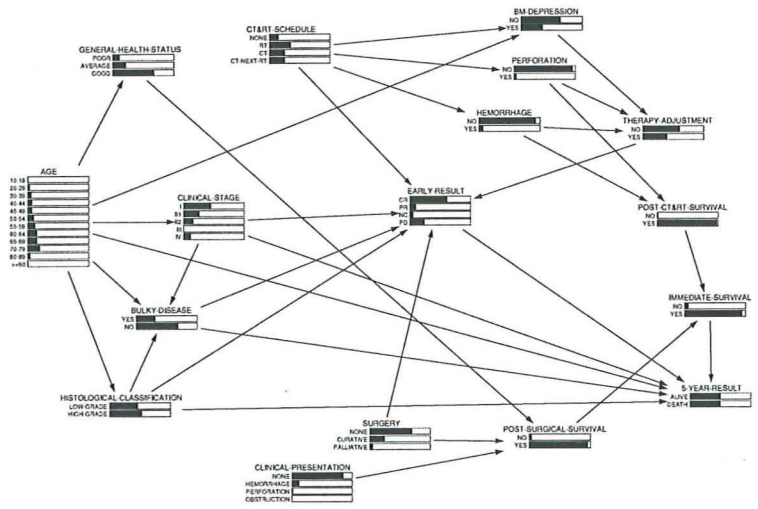

Figure 4: Declarative Bayesian network as designed with the help of expert clinical oncologists.

Fig. 3 shows how a partial model is estimated from a declarative model using equation (3) and employed to estimate the probabilities for a FAN classifier. Varying the amount of background knowledge we have at our disposal enables us to investigate the relative merits of knowledge of different degrees of completeness. The upper bound of completeness is formed by the knowledge represented in the declarative Bayesian network.

3 Non-Hodgkin lymphoma model and data

In this research, we used a Bayesian network incorporating most factors relevant for the management of the uncommon disease *gastric non-Hodgkin lymphoma* (NHL for short), referred to as the *declarative model*, which is shown in Fig. 4. It is fully based on expert knowledge and has been developed in collaboration with clinical experts from the Netherlands Cancer Institute (NKI) [6]. The model has been shown to contain a significant amount of high quality knowledge [1]. Furthermore, we are in the possession of a database containing 137 patients which have been diagnosed with gastric NHL.

Note that our use of both a high quality declarative model and an accompanying patient database is fairly uncommon, since most machine learning research is either based on the availability of large amounts of data or on a declarative model from which the data is generated. These models and data are often explicitly designated for benchmarking purposes, but it is not known and even doubted whether they properly represent the real-world situation [5]. Therefore, we have chosen to use both a model and a dataset taken directly from clinical practise. The declarative model serves as the background knowl-

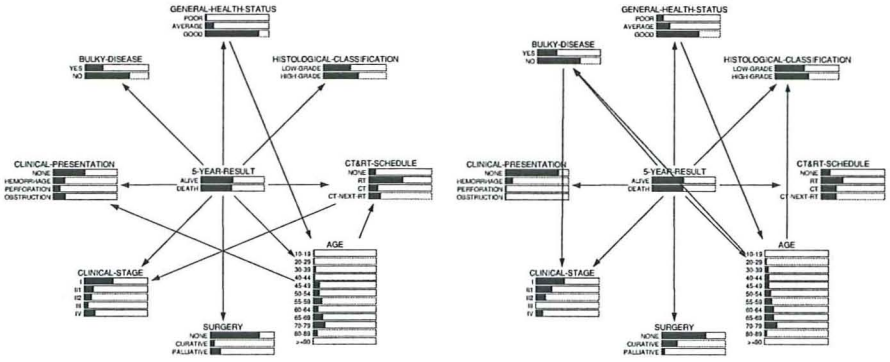

Figure 5: Differing resulting structures for data-driven FAN classifiers (left) and model-driven FAN classifiers (right) for the class-variable 5-YEAR-RESULT.

edge we have at our disposal and we will show how its exploitation may assist in the construction of Bayesian classifiers. We investigate whether the use of partial background knowledge is a feasible strategy in case of limited availability of data.

We excluded post-treatment variables and have built FAN classifiers, where the structure and underlying probability distributions are either learnt from the available patient data or estimated directly from the (partial) declarative model using equation (2). Notice that resulting models differ when structure is learnt from either patient data or the declarative model (Fig. 5).

Classifiers were evaluated by computing classification accuracy and logarithmic score for 137 patient cases for the class-variable 5-YEAR-RESULT. This variable represents whether a patient has died from NHL (DEATH) or lives (ALIVE) five years after therapy. For the classifiers learnt from patient data leave-one-out cross-validation was carried out in order to prevent overfitting artifacts. Probability distributions of the classifiers were compared with that of the declarative model by means of KL divergence.

4 Results

4.1 Building classifiers from data or background knowledge

The results for both classification accuracy and logarithmic score (Fig. 6) show that performance was consistently better for the model-driven classifiers than for the data-driven classifiers. Construction of a classifier from a database of a limited number of cases obviously leads to a performance degradation and the use of background knowledge considerably enhances classifier quality. Fig. 6

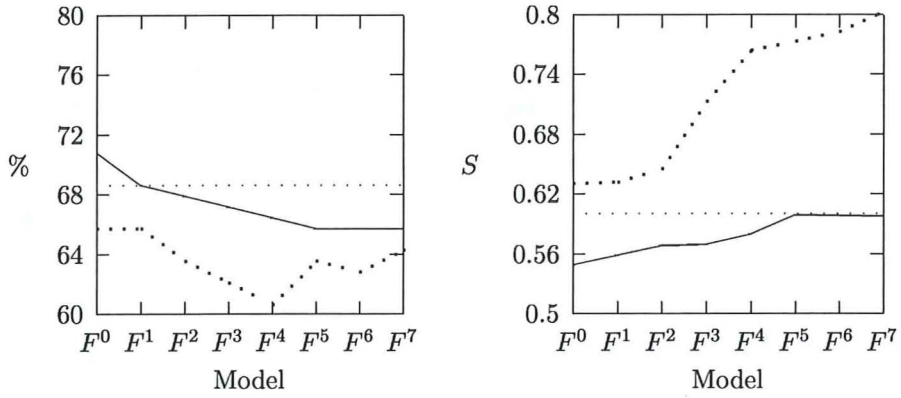

Figure 6: Classification accuracy (left) and logarithmic score (right) for Bayesian classifiers with a varying number of arcs learnt from either patient data (dotted line) or the declarative model (solid line). Classification accuracy and logarithmic score for the declarative model are shown for reference (straight line).

also shows that model-driven FAN classifiers attained better performance than the declarative model, which is task-neutral and not optimised for classification.

With regard to the naive data-driven classifier, we observed a higher logarithmic score than that of the naive model-driven classifier. Since the structures are equivalent, this must be caused by an incorrect estimation of the conditional probabilities. This is also evident from the discrepancies between the prior probabilities for classifiers built either from data or from background knowledge, as depicted in Fig. 5.

When comparing FAN classifiers we have found that entirely different dependencies were added due to large differences in CMI for variable pairs when computed either from patient data or background knowledge. The first dependency which was added in case of patient data is the dependency between CT&RT-SCHEDULE (chemotherapy and radiotherapy schedule) and CLINICAL-STAGE having a CMI of 0.212. An indirect dependency with a CMI of 0.0112 indeed exists between these variables, since the two post-treatment variables EARLY-RESULT and 5-YEAR-RESULT are mutual descendants (Fig. 4). Because post-treatment information is unknown at the time of therapy administration, clinicians tend to base therapy selection directly on the clinical stage of the tumour. This is an example of a discrepancy between expert opinion and clinical practise, which must be taken into account when validating a model based on patient data.

Performance of data-driven classifiers containing more arcs tended to decrease. This is caused by the fact that the incorrect estimation of conditional probabilities is amplified by adding more arcs. The addition of a parent with n states multiplies the number of possible parent configurations of a vertex by n.

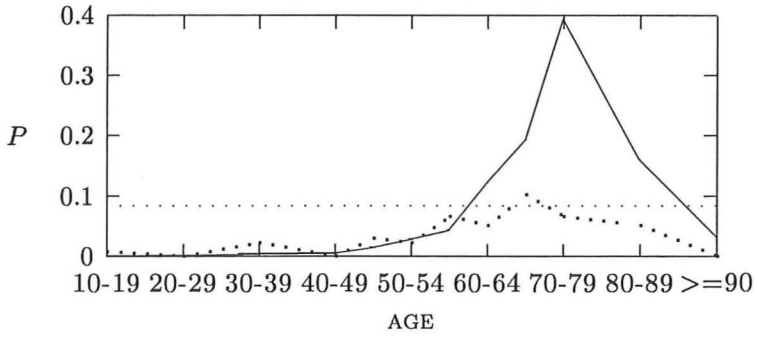

Figure 7: The probability distribution $P(\text{AGE} \mid \text{GHS=POOR}, \text{5-YEAR-RESULT=DEATH})$ is estimated as a uniform distribution since there is no data present for this configuration and the Dirichlet prior is uniform. Note that a Dirichlet prior chosen as the marginal distribution $P(\text{AGE} \mid \text{5-YEAR-RESULT=DEATH})$ computed from patient data (dotted line) comes closer to the distribution computed from the declarative model (solid line).

Table 1: Kullback-Leibler divergences.

	F^0	F^1	F^2	F^3	F^4	F^5	F^6	F^7
Model-driven	0.52	0.27	0.22	0.18	0.15	0.14	0.13	0.13
Data-driven	6.56	6.58	8.40	9.24	11.55	11.56	12.36	13.77

For instance, a large increase in logarithmic score going from model F_d^2 to F_d^3 was observed. In this case, a dependency between GHS (general health status) and AGE was added. There is however no patient data available on the age distribution when GHS takes on the value POOR, such that a uniform Dirichlet prior will be assumed, which is inconsistent with the knowledge contained in the declarative model. Fig. 7 shows an example of such inconsistencies and illustrates the benefits of using marginal probability distributions as Dirichlet priors.

This incorrect estimation is also evident from the increasing KL divergence between the declarative model and data-driven classifiers with increasing structural complexity (Table 1). Note that this decrease in performance was also observed for model-driven classifiers, in which case amplification of incorrect estimation cannot be an explanation, because conditional probabilities can be reliably estimated from the declarative model. We again expect discrepancies between expert opinion and clinical practise to play a role in this case. It is improbable that the naive classifier is simply the best representation of the dependencies within the model since KL divergence was shown to decrease for model-driven classifiers of increasing structural complexity.

In order to test whether a naive classifier always performs best for this domain, we have generated a random sample of 1370 cases from the declarative model by means of *probabilistic logic sampling* [4]. When validating the model based on this sample we found that logarithmic score decreased from 0.545 for the naive model to 0.523 for the TAN model. Thus, TAN models are in principle able to perform better than a naive model, but for this domain, improvement is only marginal. The reason for this marginal improvement is explained as follows.

When comparing the CMI between variables computed from either background knowledge or patient data, we have found that there is only one dependency between GHS (general health status) and AGE showing a high CMI of 0.173 when computed from background knowledge, whereas there are many such combinations when computed from patient data. Let P_0 and P_1 denote the probability distributions for model F_m^0 and model F_m^1 encoding this dependency. The differences in classification performance for these models are then specified by $P_0(\text{5-YEAR-RESULT} \mid \text{AGE}, \text{GHS})$ and $P_1(\text{5-YEAR-RESULT} \mid \text{AGE}, \text{GHS})$ which can be computed from

$$\frac{P(\text{AGE} \mid \text{GHS},\text{5-YEAR-RESULT})}{P(\text{AGE} \mid \text{GHS})} \frac{P(\text{GHS} \mid \text{5-YEAR-RESULT})P(\text{5-YEAR-RESULT})}{P(\text{GHS})},$$

where the last component is constant for both F_m^0 and F_m^1 and the first component reduces to $P_0(\text{AGE} \mid \text{5-YEAR-RESULT})/P_0(\text{AGE})$ for model F_m^0. When we compare $\delta(P_1(\text{AGE} \mid \text{GHS},\text{5-YEAR-RESULT}), P_0(\text{AGE} \mid \text{GHS}, \text{5-YEAR-RESULT}))$ and $\delta(P_1(\text{5-YEAR-RESULT} \mid \text{AGE}, \text{GHS}), P_0(\text{5-YEAR-RESULT} \mid \text{AGE}, \text{GHS}))$ we find Kullback-Leibler divergences of respectively 2.00 and 0.135.

Let $c^*_{\{\text{AGE,GHS}\}}$ denote the value of the class variable 5-YEAR-RESULT for evidence $\{\text{AGE}, \text{GHS}\}$, classified using $P_1(\text{5-YEAR-RESULT} \mid \text{AGE}, \text{GHS})$. The difference between the logarithmic score $S_{\{\text{AGE,GHS}\}}$ of F_m^0 and F_m^1 for evidence $\{\text{AGE}, \text{GHS}\}$ can be written as

$$P_1(c^*_{\{\text{AGE,GHS}\}} \mid \text{AGE}, \text{GHS}) \log \frac{P_1(c^*_{\{\text{AGE,GHS}\}} \mid \text{AGE}, \text{GHS})}{P_0(c^*_{\{\text{AGE,GHS}\}} \mid \text{AGE})},$$

and the KL divergence between models F_m^0 and F_m^1 can be written as

$$\sum_{\text{AGE,GHS,5-YEAR-RESULT}} P_1(\text{AGE}, \text{GHS}, \text{5-YEAR-RESULT}) \log \frac{P_1(\text{AGE} \mid \text{GHS}, \text{5-YEAR-RESULT})}{P_0(\text{AGE} \mid \text{5-YEAR-RESULT})}.$$

There is little impact on logarithmic score since this is dependent on factors $P(\text{5-YEAR-RESULT} \mid \text{AGE}, \text{GHS})$, which show only little KL divergence between models F_m^0 and F_m^1. Impact on KL divergence between models F_m^0 and F_m^1 is high, since this is dependent on factors $P(\text{AGE} \mid \text{GHS}, \text{5-YEAR-RESULT})$.

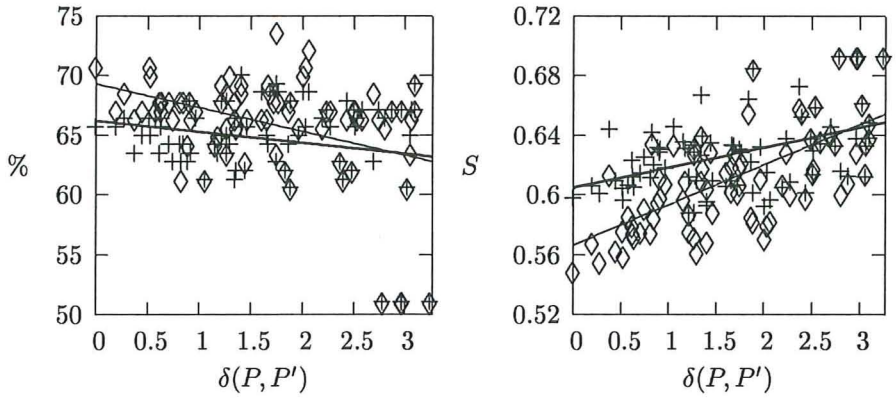

Figure 8: Regression results on classification accuracy and logarithmic score for the naive classifier F_m^0 ($\diamond$, thin line) and TAN classifier F_m^7 ($+$, thick line) for partial models containing varying amounts of partial background knowledge as measured by the KL divergence between the declarative model $\mathcal{B} = (G, P)$ and partial models $\mathcal{B}' = (G', P')$.

4.2 Partial background knowledge and Bayesian classification

Although the benefit of using background knowledge has been demonstrated in the previous sections, it will not usually be the case that full knowledge of the domain is available. Instead, one expects the expert to deliver partial knowledge about the structure and underlying probabilities of the domain. In this section we investigate how partial specifications influence the quality of Bayesian classifiers. To this end, we created partial models retaining 0, 5, 10, 15, 20, 25 and all 32 arcs of the original declarative model. In total 77 different partial models were generated and the KL divergence between the declarative and partial models was computed. From these models we have generated model-driven FAN classifiers F_m^0 and F_m^7. Linear regressions on classification accuracy and logarithmic score for these models are shown in Fig. 8.

Results show that on average, performance increases as $\delta(P, P')$ becomes smaller. This demonstrates that the use of partial background knowledge is indeed a feasible alternative to the use of data for the construction of Bayesian classifiers. Performance of naive and TAN classifiers coincides when relevant dependencies in the partial model can be fully represented within the conditional probability tables of the naive classifier. The outliers were identified to be partial models where the class-variable 5-YEAR-RESULT is a disconnected vertex.

On average, a partial model containing 10 arcs attained a performance similar to that of the model which was learnt from data. The benefits of using more fine-grained background knowledge are apparent, even for naive

classifiers, for which conditional probabilities $P(E_i \mid C)$ have to be estimated. As the variables E_i and C may be located far apart within the declarative model, more complete knowledge will increase the accuracy of estimating these probabilities.

Irrelevant vertices are not taken into account when constructing partial models. Hence, arcs may be removed which do not influence the quality of the background knowledge represented in the model with respect to the classification task. On the other hand, naively removing arcs from the declarative model may disconnect the class-variable from the rest of the model, reducing model quality severely. In practise, one expects a domain expert to provide a partial model which expresses knowledge relevant to the classification task.

5 Conclusion

Many real-world problems are characterised by the absence of sufficient statistical data about the domain. Most algorithms for constructing Bayesian classifiers are highly data-driven and therefore incapable of producing acceptable results in such data-poor domains. In this article we have formalised the notion of partial background knowledge and introduced the concept of a partial model. We presented a method for constructing model-driven classifiers from partial background knowledge and showed that they outperform data-driven classifiers for data-poor domains. This even holds for the naive classifier, which is highly biased but shows enough variance to encode at least some dependence information.

The use of both a model and a dataset taken directly from clinical practise enabled us to show that when the structural complexity of data-driven classifiers is increased, performance can be considerably reduced due to the amplification of incorrect estimation of conditional probabilities. More importantly, we have shown that for model-driven classifiers differences between expert opinion and clinical practise are likely to be the major source of this decrease. Such interactions between model and data are absent when classifiers are learnt or validated by artificial models or datasets.

A comparison of logarithmic score and KL divergence demonstrated that even though more complex classifier structures encode very different conditional probability distributions, this may exert only a marginal positive influence on logarithmic score and consequently classification accuracy. This marginal performance gain depends heavily on the availability of large amounts of high quality information, which is often not the case. We expect performance for structurally more complex classifiers to improve only when the CMIs for the added dependencies are high and if we have enough data at our disposal to warrant these dependencies.

We have demonstrated that for a real-world problem, background knowledge offers a significant contribution to improving the quality of learnt classifiers and even becomes invaluable since available data is often noisy, small and incomplete. Note that our operationalization of partial background knowledge is

only one of the many forms of background knowledge one may wish to include. In a real-world setting, a proper mix should be determined in terms of the use of various kinds of background knowledge on one hand and learning based on data on the other hand. As Mitchell has already remarked in the 1980's: "If bias and initial knowledge are at the heart of the ability to generalize beyond observed data, then efforts to study machine learning must focus on the *combined* use of prior knowledge, biases and observation in guiding the learning process" [7]. The development of techniques for using background knowledge in order to improve the quality of Bayesian networks will be the focus of our future research.

References

[1] C. Bielza, J. A. Fernández del Pozo, and P. J. F. Lucas. Finding and explaining optimal treatments. In *AIME 2003*, pages 299–303, 2003.

[2] R.G. Cowell, A.P Dawid, and D. Spiegelhalter. Sequential model criticism in probabilistic expert systems. *PAMI*, 15(3):209–219, 1993.

[3] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[4] M. Henrion. Propagation of uncertainty by probabilistic logic sampling in Bayes' networks. In *Proceedings of Uncertainty in Artificial Intelligence*, volume 2, pages 149–163, 1988.

[5] P.J.F. Lucas. Restricted Bayesian network structure learning. In J.A. Gâmez, S. Moral, and A. Salmeron, editors, *Advances in Bayesian Networks, Studies in Fuzziness and Soft Computing*, volume 146, pages 217–232. Springer-Verlag, Berlin, 2004.

[6] P.J.F. Lucas, H. Boot, and B.G. Taal. Computer-based decision support in the management of primary gastric non-Hodgkin lymphoma. *Methods of Information in Medicine*, 37:206–219, 1998.

[7] T.M. Mitchell. The need for biases in learning generalizations. Technical report, Rutgers University, Department of Computer Science, 1980.

[8] M. Pazzani. Searching for dependencies in Bayesian classifiers. In *Learning from data: Artificial intelligence and statistics V*, pages 239–248. New York, NY: Springer-Verlag, 1996.

[9] S.L. Lauritzen, A.P. Dawid, B.N. Larsen, and H.G. Leimer. Independence properties of directed Markov fields. *Networks*, 20:491–506, 1990.